Statistical Methods for Analysis with Missing Data

Lecture 6: the EM algorithm

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lecture

Introduction to likelihood-based approaches to handling missing data:

- In general, we need to posit parametric models for the study variables and for the response mechanism
- Response mechanism can be ignored if MAR + separability of parameters

$$L_{obs}(\theta,\psi) \overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)},\psi)\right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta)\, dz_{i(\bar{r}_i)}\right]}_{\text{Provides MLE of } \theta \text{ under MAR}}$$

- Finding the MLE of θ might be complicated, even under ignorability

Today's Lecture

The Expectation-Maximization (EM) algorithm

- General derivation
- Simplification under MAR
- Simplification under exponential families
- Monte Carlo EM

Reading

- Pages 62 to 75, Chapter 3 of Davidian and Tsiatis (not required)
- Sean Borman's online tutorial on the EM algorithm (recommended): `https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf`

The Expectation-Maximization Algorithm

- The EM algorithm was presented formally by Dempster, Laird & Rubin (JRSSB, 1977), although similar ideas had appeared earlier

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

Harvard University and Educational Testing Service

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

Keywords: MAXIMUM LIKELIHOOD; INCOMPLETE DATA; EM ALGORITHM; POSTERIOR MODE

- A general scheme for deriving maximization algorithms when the likelihood can be expanded in terms of latent or missing variables

Derivation of the EM Algorithm

- We need to maximize *observed-data likelihood*, with generic term given by

$$\ell_{obs}(\vartheta) = \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\ dz_{(\bar{r})}$$

  where $\vartheta = (\theta, \psi)$ and $p(r, z \mid \vartheta) = p(r \mid z, \psi)p(z \mid \theta)$

- Goal: construct sequence $\vartheta^{(0)}, \vartheta^{(1)}, \ldots$ that converges to the MLE $\hat{\vartheta}$

- Remember: maximizing $\ell_{obs}(\vartheta)$ is equivalent to maximizing $\log \ell_{obs}(\vartheta)$

Fun Fact: Jensen's Inequality

If f is a concave function, and X is a random variable, then

$$f[E(X)] \geq E[f(X)]$$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

$$\log \ell_{obs}(\vartheta) = \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\, dz_{(\bar{r})}$$

$$= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})}$$

$$= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})}$$

$$\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] dz_{(\bar{r})}$$

$$\equiv h(\vartheta \mid \vartheta^{(t)})$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

$$\log \ell_{obs}(\vartheta) = \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\, dz_{(\bar{r})}$$

$$= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})}$$

$$= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})}$$

$$\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] dz_{(\bar{r})}$$

$$\equiv h(\vartheta \mid \vartheta^{(t)})$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

$$\begin{aligned}
\log \ell_{obs}(\vartheta) &= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\, dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})} \\
&\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] dz_{(\bar{r})} \\
&\equiv h(\vartheta \mid \vartheta^{(t)})
\end{aligned}$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

$$\begin{aligned}
\log \ell_{obs}(\vartheta) &= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\ dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\ dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\ dz_{(\bar{r})} \\
&\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right]\ dz_{(\bar{r})} \\
&\equiv h(\vartheta \mid \vartheta^{(t)})
\end{aligned}$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

$$
\begin{aligned}
\log \ell_{obs}(\vartheta) &= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\ dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\ dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\ dz_{(\bar{r})} \\
&\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right]\ dz_{(\bar{r})} \\
&\equiv h(\vartheta \mid \vartheta^{(t)})
\end{aligned}
$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

Say we have $\vartheta^{(t)}$, note that

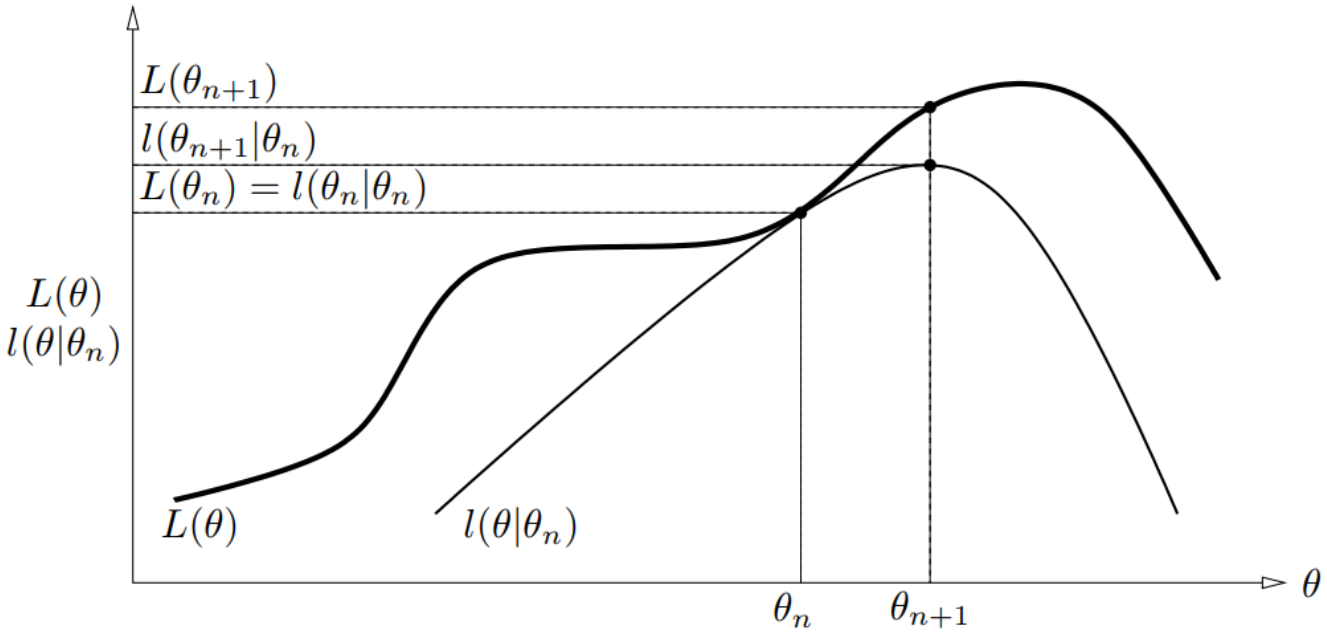
$$
\begin{aligned}
\log \ell_{obs}(\vartheta) &= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta)\, dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(r, z \mid \vartheta) \frac{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})} \\
&= \log \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})}\, dz_{(\bar{r})} \\
&\overset{\text{Jensen's}}{\geq} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] dz_{(\bar{r})} \\
&\equiv h(\vartheta \mid \vartheta^{(t)})
\end{aligned}
$$

HW2: show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$

Derivation of the EM Algorithm

- Since
  - $\log \ell_{obs}(\vartheta) \geq h(\vartheta \mid \vartheta^{(t)})$
  - $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$
- A value $\vartheta^{(t+1)}$ that maximizes $h(\vartheta \mid \vartheta^{(t)})$ will increase the value of $\log \ell_{obs}(\vartheta)$
- Idea: iteratively maximize $h(\vartheta \mid \vartheta^{(t)})$

EM in a Picture



Taken from Sean Borman's online tutorial

Derivation of the EM Algorithm

$$\vartheta^{(t+1)} = \arg\max_{\vartheta} h(\vartheta \mid \vartheta^{(t)})$$

$$= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] dz_{(\bar{r})}$$

$$= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z \mid \vartheta) \, dz_{(\bar{r})}$$

$$= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z_{(r)}, z_{(\bar{r})} \mid \vartheta) \, dz_{(\bar{r})}$$

$$= \arg\max_{\vartheta} E\left[\log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta) \mid R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)} \right]$$

Derivation of the EM Algorithm

$$
\begin{aligned}
\vartheta^{(t+1)} &= \arg\max_{\vartheta} \; h(\vartheta \mid \vartheta^{(t)}) \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z_{(r)}, z_{(\bar{r})} \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} E \left[\log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta) \;\middle|\; R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)} \right]
\end{aligned}
$$

Derivation of the EM Algorithm

$$
\begin{aligned}
\vartheta^{(t+1)} &= \arg\max_{\vartheta} \; h(\vartheta \mid \vartheta^{(t)}) \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z_{(r)}, z_{(\bar{r})} \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} E \left[\log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta) \;\middle|\; R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)} \right]
\end{aligned}
$$

Derivation of the EM Algorithm

$$
\begin{aligned}
\vartheta^{(t+1)} &= \arg\max_{\vartheta} \; h(\vartheta \mid \vartheta^{(t)}) \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z_{(r)}, z_{(\bar{r})} \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} E \left[\log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta) \;\middle|\; R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)} \right]
\end{aligned}
$$

Derivation of the EM Algorithm

$$
\begin{aligned}
\vartheta^{(t+1)} &= \arg\max_{\vartheta} \; h(\vartheta \mid \vartheta^{(t)}) \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log \left[\frac{p(r, z \mid \vartheta)}{p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})} \right] \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) \log p(r, z_{(r)}, z_{(\bar{r})} \mid \vartheta) \; dz_{(\bar{r})} \\
&= \arg\max_{\vartheta} \; E \left[\; \log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta) \;\; \middle| \;\; R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)} \right]
\end{aligned}
$$

The EM Algorithm

- The Expectation step:

$$Q(\vartheta \mid \vartheta^{(t)}) = E\ [\ \log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta)\ \mid\ R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)}]$$

- The Maximization step:

$$\vartheta^{(t+1)} = \arg\max_{\vartheta}\ Q(\vartheta \mid \vartheta^{(t)})$$

- The algorithm is run until some convergence criterion is satisfied

The Expectation Step

$$
\begin{aligned}
Q(\vartheta \mid \vartheta^{(t)}) &= E\ [\ \log p(r, z_{(r)}, Z_{(\bar{r})} \mid \vartheta)\ \mid\ R = r, Z_{(r)} = z_{(r)}, \vartheta^{(t)}] \\
&= \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})[\ \log p(r \mid z, \psi) + \log p(z \mid \theta)\ ]\ dz_{(\bar{r})} \\
&= Q_{\psi}(\psi \mid \vartheta^{(t)}) + Q_{\theta}(\theta \mid \vartheta^{(t)}),
\end{aligned}
$$

where

$$
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}) = \frac{p(r, z \mid \vartheta^{(t)})}{p(r, z_{(r)} \mid \vartheta^{(t)})},
$$

which, generally, is not nice-looking!

The Maximization Step

$\vartheta^{(t+1)} = (\psi^{(t+1)}, \theta^{(t+1)})$, where

$$\psi^{(t+1)} = \arg\max_{\psi} Q_{\psi}(\psi \mid \vartheta^{(t)})$$

$$\theta^{(t+1)} = \arg\max_{\theta} Q_{\theta}(\theta \mid \vartheta^{(t)})$$

The EM Algorithm Under MAR

Again, life is easier under MAR

$$
\begin{aligned}
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta) &= \frac{p(r, z \mid \vartheta)}{p(r, z_{(r)} \mid \vartheta)} \\
&= \frac{p(r \mid z, \psi) p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}} \\
&\overset{\text{MAR}}{=} \frac{p(r \mid z_{(r)}, \psi) p(z \mid \theta)}{p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= \frac{p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= p(z_{(\bar{r})} \mid z_{(r)}, \theta)
\end{aligned}
$$

The EM Algorithm Under MAR

Again, life is easier under MAR

$$
\begin{aligned}
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta) &= \frac{p(r, z \mid \vartheta)}{p(r, z_{(r)} \mid \vartheta)} \\
&= \frac{p(r \mid z, \psi) p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}} \\
&\overset{\text{MAR}}{=} \frac{p(r \mid z_{(r)}, \psi) p(z \mid \theta)}{p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= \frac{p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= p(z_{(\bar{r})} \mid z_{(r)}, \theta)
\end{aligned}
$$

The EM Algorithm Under MAR

Again, life is easier under MAR

$$
\begin{aligned}
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta) &= \frac{p(r, z \mid \vartheta)}{p(r, z_{(r)} \mid \vartheta)} \\
&= \frac{p(r \mid z, \psi) p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}} \\
&\overset{\text{MAR}}{=} \frac{p(r \mid z_{(r)}, \psi) p(z \mid \theta)}{p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= \frac{p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= p(z_{(\bar{r})} \mid z_{(r)}, \theta)
\end{aligned}
$$

The EM Algorithm Under MAR

Again, life is easier under MAR

$$
\begin{aligned}
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta) &= \frac{p(r, z \mid \vartheta)}{p(r, z_{(r)} \mid \vartheta)} \\
&= \frac{p(r \mid z, \psi) p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}} \\
&\overset{\text{MAR}}{=} \frac{p(r \mid z_{(r)}, \psi) p(z \mid \theta)}{p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= \frac{p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= p(z_{(\bar{r})} \mid z_{(r)}, \theta)
\end{aligned}
$$

The EM Algorithm Under MAR

Again, life is easier under MAR

$$
\begin{aligned}
p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta) &= \frac{p(r, z \mid \vartheta)}{p(r, z_{(r)} \mid \vartheta)} \\
&= \frac{p(r \mid z, \psi) p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}} \\
&\overset{\text{MAR}}{=} \frac{p(r \mid z_{(r)}, \psi) p(z \mid \theta)}{p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= \frac{p(z \mid \theta)}{\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \, dz_{(\bar{r})}} \\
&= p(z_{(\bar{r})} \mid z_{(r)}, \theta)
\end{aligned}
$$

The EM Algorithm Under MAR

- The Expectation step:

$$Q_\theta(\theta \mid \theta^{(t)}) = E\ [\ \log p(z_{(r)}, Z_{(\bar{r})} \mid \theta)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}]$$
$$= \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid z_{(r)}, \theta^{(t)}) \log p(z \mid \theta)\ dz_{(\bar{r})}$$

- The Maximization step:

$$\theta^{(t+1)} = \arg\max_\theta\ Q_\theta(\theta \mid \theta^{(t)})$$

The EM Algorithm Under MAR

- $p(z_{(\bar{r})} \mid z_{(r)}, \theta^{(t)})$ has a nice form under some parametric models, such as multivariate normals and multinomials

- $\log p(z \mid \theta)$ decomposes nicely into expectations of sufficient statistics if $\log p(z \mid \theta)$ is in exponential-family form

The EM Algorithm Under MAR

- $p(z_{(\bar{r})} \mid z_{(r)}, \theta^{(t)})$ has a nice form under some parametric models, such as multivariate normals and multinomials

- $\log p(z \mid \theta)$ decomposes nicely into expectations of sufficient statistics if $\log p(z \mid \theta)$ is in exponential-family form

The EM Algorithm With Exponential Families

Say $p(z \mid \theta)$ belongs to an exponential family with $\theta = (\theta_1, \ldots, \theta_d)$, that is,

$$p(z \mid \theta) = b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] / c(\theta)$$

with $c(\theta) = \int b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] dz$

HW2:

- E step:

$$Q_\theta(\theta \mid \theta^{(t)}) = \sum_{s=1}^{d} \eta_s(\theta) E\left[\, T_s(z) \;\middle|\; Z_{(r)} = z_{(r)}, \theta^{(t)}\right] - \log c(\theta)$$

- M step: find $\theta^{(t+1)}$ as the solution to

$$E\left[\, T_s(Z) \;\middle|\; Z_{(r)} = z_{(r)}, \theta^{(t)}\right] = E\left[\, T_s(Z) \;\middle|\; \theta\right],\ s = 1, \ldots, d$$

The EM Algorithm With Exponential Families

Say $p(z \mid \theta)$ belongs to an exponential family with $\theta = (\theta_1, \ldots, \theta_d)$, that is,

$$p(z \mid \theta) = b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] / c(\theta)$$

with $c(\theta) = \int b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] dz$

HW2:

- E step:

$$Q_\theta(\theta \mid \theta^{(t)}) = \sum_{s=1}^{d} \eta_s(\theta) E\left[\, T_s(z) \;\middle|\; Z_{(r)} = z_{(r)}, \theta^{(t)}\right] - \log c(\theta)$$

- M step: find $\theta^{(t+1)}$ as the solution to

$$E\left[\, T_s(Z) \;\middle|\; Z_{(r)} = z_{(r)}, \theta^{(t)}\right] = E\left[\, T_s(Z) \;\middle|\; \theta\right],\ s = 1, \ldots, d$$

The EM Algorithm With Exponential Families

Say $p(z \mid \theta)$ belongs to an exponential family with $\theta = (\theta_1, \ldots, \theta_d)$, that is,

$$p(z \mid \theta) = b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] / c(\theta)$$

with $c(\theta) = \int b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] dz$

HW2:

- E step:

$$Q_\theta(\theta \mid \theta^{(t)}) = \sum_{s=1}^{d} \eta_s(\theta) E\left[\ T_s(z)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}\right] - \log c(\theta)$$

- M step: find $\theta^{(t+1)}$ as the solution to

$$E\left[\ T_s(Z)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}\right] = E\left[\ T_s(Z)\ \mid\ \theta\right],\ s = 1, \ldots, d$$

Example of EM Algorithm Under MAR

HW2:

- Let $Z = (Z_1, Z_2)$, $Z_1, Z_2 \in \{0, 1\}$
- Let $W = (W_{00}, W_{01}, W_{10}, W_{11})$, where
$$W_{kl} = I(Z_1 = k, Z_2 = l), \quad k, l \in 0, 1$$
- We can therefore write
$$p(Z_1 = z_1, Z_2 = z_2 \mid \theta) = \pi_{z_1 z_2},$$
or
$$p[W = (w_{00}, w_{01}, w_{10}, w_{11}) \mid \theta] = \prod_{k,l} \pi_{kl}^{w_{kl}},$$
where $\theta = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$
- What is $p(Z_1 = z_1 \mid Z_2 = z_2, \theta)$?
- Derive
$$Q_\theta(\theta \mid \theta^{(t)}) = E\ [\ \log p(z_{(r)}, Z_{(\bar{r})} \mid \theta)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}]$$
$$= \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid z_{(r)}, \theta^{(t)}) \log p(z \mid \theta)\ dz_{(\bar{r})}$$

Example of EM Algorithm Under MAR

HW2:

- Let $Z_i = (Z_{i1}, Z_{i2})$, $Z_{i1}, Z_{i2} \in \{0, 1\}$, Z_i's are i.i.d.
- Let $R_i = (R_{i1}, R_{i2})$, $R_{i1}, R_{i2} \in \{0, 1\}$, R_i's are i.i.d.
- Let $W_i = (W_{i00}, W_{i01}, W_{i10}, W_{i11})$, where

$$W_{ikl} = I(Z_{i1} = k, Z_{i2} = l), \quad k, l \in 0, 1$$

- We can therefore write

$$L(\theta) = \prod_{i,k,l} \pi_{kl}^{w_{ikl}},$$

  where $\theta = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$

- Derive

$$Q_\theta(\theta \mid \theta^{(t)})$$

- Derive

$$\theta^{(t+1)} = \arg\max_\theta \ Q_\theta(\theta \mid \theta^{(t)})$$

Comments

- Under regularity conditions, the EM algorithm is guaranteed to converge to a stationary point of the observed-data log-likelihood; however, this could be a local, global, or a saddle point
- Jeff Wu (1983) showed that the EM algorithm will converge if the objective function has a unique maximum in the interior of the parameter space
- In practice, it is common to run the algorithm starting from different starting points
- The EM algorithm converges linearly, whereas other optimization methods such as Newton-Raphson converge quadratically
- The EM algorithm doesn't provide standard error estimates

Comments

- Under regularity conditions, the EM algorithm is guaranteed to converge to a stationary point of the observed-data log-likelihood; however, this could be a local, global, or a saddle point
- Jeff Wu (1983) showed that the EM algorithm will converge if the objective function has a unique maximum in the interior of the parameter space
- In practice, it is common to run the algorithm starting from different starting points
- The EM algorithm converges linearly, whereas other optimization methods such as Newton-Raphson converge quadratically
- The EM algorithm doesn't provide standard error estimates

Comments

- Under regularity conditions, the EM algorithm is guaranteed to converge to a stationary point of the observed-data log-likelihood; however, this could be a local, global, or a saddle point
- Jeff Wu (1983) showed that the EM algorithm will converge if the objective function has a unique maximum in the interior of the parameter space
- In practice, it is common to run the algorithm starting from different starting points
- The EM algorithm converges linearly, whereas other optimization methods such as Newton-Raphson converge quadratically
- The EM algorithm doesn't provide standard error estimates

Comments

- Under regularity conditions, the EM algorithm is guaranteed to converge to a stationary point of the observed-data log-likelihood; however, this could be a local, global, or a saddle point
- Jeff Wu (1983) showed that the EM algorithm will converge if the objective function has a unique maximum in the interior of the parameter space
- In practice, it is common to run the algorithm starting from different starting points
- The EM algorithm converges linearly, whereas other optimization methods such as Newton-Raphson converge quadratically
- The EM algorithm doesn't provide standard error estimates

Comments

- Under regularity conditions, the EM algorithm is guaranteed to converge to a stationary point of the observed-data log-likelihood; however, this could be a local, global, or a saddle point
- Jeff Wu (1983) showed that the EM algorithm will converge if the objective function has a unique maximum in the interior of the parameter space
- In practice, it is common to run the algorithm starting from different starting points
- The EM algorithm converges linearly, whereas other optimization methods such as Newton-Raphson converge quadratically
- The EM algorithm doesn't provide standard error estimates

Monte Carlo EM Algorithm

- For complex models, we might not be able to compute a closed form for

$$Q(\vartheta \mid \vartheta^{(t)}) = \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})[\ \log p(r \mid z, \psi) + \log p(z \mid \theta)\ ]\ dz_{(\bar{r})}$$

- The idea: use Monte Carlo integration!
  - Draw $z_{(\bar{r})}^{(1)}, \ldots, z_{(\bar{r})}^{(M)}$ from the conditional distr. of $Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}$
  - Approximate $Q(\vartheta \mid \vartheta^{(t)})$ as

$$\frac{1}{M} \sum_{m=1}^{M} \left[\log p(r \mid z_{(r)}, z_{(\bar{r})}^{(m)}, \psi) + \log p(z_{(r)}, z_{(\bar{r})}^{(m)} \mid \theta) \right]$$

  - Needs to be done at each step of EM, for each i*th* sample point!
  - Can be very computationally intensive
  - Monte Carlo error ruins convergence guarantees of the EM

- My point of view: if I can sample from $p(Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta)$, I might as well just take a full Bayesian approach! (we'll see this later)

Monte Carlo EM Algorithm

- For complex models, we might not be able to compute a closed form for

$$Q(\vartheta \mid \vartheta^{(t)}) = \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})[\ \log p(r \mid z, \psi) + \log p(z \mid \theta)\ ]\ dz_{(\bar{r})}$$

- The idea: use Monte Carlo integration!
  - Draw $z_{(\bar{r})}^{(1)}, \ldots, z_{(\bar{r})}^{(M)}$ from the conditional distr. of $Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}$
  - Approximate $Q(\vartheta \mid \vartheta^{(t)})$ as

$$\frac{1}{M} \sum_{m=1}^{M} \left[\log p(r \mid z_{(r)}, z_{(\bar{r})}^{(m)}, \psi) + \log p(z_{(r)}, z_{(\bar{r})}^{(m)} \mid \theta) \right]$$

  - Needs to be done at each step of EM, for each *ith* sample point!
  - Can be very computationally intensive
  - Monte Carlo error ruins convergence guarantees of the EM

- My point of view: if I can sample from $p(Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta)$, I might as well just take a full Bayesian approach! (we'll see this later)

Monte Carlo EM Algorithm

- For complex models, we might not be able to compute a closed form for

$$Q(\vartheta \mid \vartheta^{(t)}) = \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})[\ \log p(r \mid z, \psi) + \log p(z \mid \theta)\ ]\ dz_{(\bar{r})}$$

- The idea: use Monte Carlo integration!
  - Draw $z_{(\bar{r})}^{(1)}, \ldots, z_{(\bar{r})}^{(M)}$ from the conditional distr. of $Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}$
  - Approximate $Q(\vartheta \mid \vartheta^{(t)})$ as

$$\frac{1}{M} \sum_{m=1}^{M} \left[\log p(r \mid z_{(r)}, z_{(\bar{r})}^{(m)}, \psi) + \log p(z_{(r)}, z_{(\bar{r})}^{(m)} \mid \theta) \right]$$

  - Needs to be done at each step of EM, for each i*th* sample point!
  - Can be very computationally intensive
  - Monte Carlo error ruins convergence guarantees of the EM

- My point of view: if I can sample from $p(Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta)$, I might as well just take a full Bayesian approach! (we'll see this later)

Monte Carlo EM Algorithm

- For complex models, we might not be able to compute a closed form for

$$Q(\vartheta \mid \vartheta^{(t)}) = \int_{\mathcal{Z}_{(\bar{r})}} p(z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)})[\ \log p(r \mid z, \psi) + \log p(z \mid \theta)\ ]\ dz_{(\bar{r})}$$

- The idea: use Monte Carlo integration!
  - Draw $z_{(\bar{r})}^{(1)}, \ldots, z_{(\bar{r})}^{(M)}$ from the conditional distr. of $Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta^{(t)}$
  - Approximate $Q(\vartheta \mid \vartheta^{(t)})$ as

$$\frac{1}{M} \sum_{m=1}^{M} \left[\log p(r \mid z_{(r)}, z_{(\bar{r})}^{(m)}, \psi) + \log p(z_{(r)}, z_{(\bar{r})}^{(m)} \mid \theta) \right]$$

  - Needs to be done at each step of EM, for each i*th* sample point!
  - Can be very computationally intensive
  - Monte Carlo error ruins convergence guarantees of the EM

- My point of view: if I can sample from $p(Z_{(\bar{r})} \mid r, z_{(r)}, \vartheta)$, I might as well just take a full Bayesian approach! (we'll see this later)

The EM Algorithm under MAR

For standard errors / variance-covariance matrices

- Tom Louis (1982, JRSSB)
- Supplemental EM algorithm of Meng & Rubin (1991, JASA)
- Efron's Bootstrap

Summary

Main take-aways from today's lecture:

- EM algorithm can be convenient in some classes of parametric models

Next lecture:

- Implementations of EM algorithms in R
- Standard errors: Bootstrap + EM

Summary

Main take-aways from today's lecture:

- EM algorithm can be convenient in some classes of parametric models

Next lecture:

- Implementations of EM algorithms in R
- Standard errors: Bootstrap + EM